

A Multi-Model Approach to Analysis of Environmental Phenomena

O. Giustolisi^a, A. Doglioni^a, D.A. Savic^b and B.W. Webb^c

^a *Civil and Environmental Engineering Department, Engineering Faculty of Taranto, Technical University of Bari, Italy - o.giustolisi@poliba.it*

^b *Centre for Water Systems, Department of Engineering, School of Engineering, Computer Science and Mathematics, University of Exeter, United Kingdom*

^c *Department of Geography, School of Geography, Archaeology & Earth Resources, University of Exeter, United Kingdom*

Abstract: This paper introduces a novel data-driven methodology named Evolutionary Polynomial Regression (EPR), which permits the multi-purpose modelling of physical phenomena, through the simultaneous solution of a number of models. Multipurpose modelling or “multi-modelling”, enables the user to make a more robust choice of those models aimed at (a) the knowledge based on data modelling, (b) on-line and off-line forecasting, and (c) data augmentation (i.e. infilling of missing data in time series). This methodology is particularly useful in modelling environmental phenomena, for which it is usually impossible to obtain physical data at a laboratory scale. In particular, the non-linearity of phenomena and non Gaussian nature of background noise make on-line forecasting complex, and where data are available, they often contain discontinuities (i.e. missing data). The use of EPR in modelling and analysis is illustrated by application to a case study containing all these limitations. The application of EPR to thermal behaviour of a stream gives not only a good physical insight of the phenomenon, but also allows infilling of missing data, resulting in good models that forecast the water temperature.

Keywords: Data reconstruction; Knowledge discovery in data; Environmental modelling; Evolutionary computing; Evolutionary Polynomial Regression.

1. INTRODUCTION

Modelling of environmental phenomena usually relies on sampled data, which are often incomplete. Ideally, analysis should provide new insights into the phenomena, give accurate forecasting of the output for a range of inputs and outputs and fill in gaps in data records. This can be achieved by creating a range of specific models, i.e. models chosen for well-defined purposes, although the construction and choice of the models is often challenging. Environmental phenomena are typically non-linear in their dynamics and affected by non Gaussian background noise. In the models, these effects must be reproduced as accurately as possible. The temptation is to use complex non-linear modelling strategies, to better describe the phenomena. Unfortunately, the answers

from these are very difficult to interpret from a physical aspect.

An additional problem relates to discontinuities, i.e. gaps, often present in data records. On the one hand, we are interested in “reconstructing” that information contained in missing data, without losing the physics of the phenomenon. On the other hand, we do not know how to choose the best model for this purpose, because we have no data to define a traditional performance indicator.

This paper presents the Evolutionary Polynomial Regression (EPR) technique a novel, model-based reconstruction technique capable of reconstructing data series containing information about the physical phenomena [Giustolisi et al., 2004a]. It also provides simple well defined effective models useful both for on-line forecasting and simulation.

Such models usually are simple polynomial structures where each monomial can contain user-defined functions. These structures can improve physical interpretation of the studied phenomenon too [Giustolisi et al., 2004b]. EPR has the advantage of combining evolutionary algorithms with traditional numerical regression [Giustolisi and Savic, 2004a]. EPR is an incremental development of a hybrid methodology [Davidson et al. 1999; 2003] which incorporated least squares optimization within symbolic regression.

Thus, EPR is a hybrid system capable of producing a series of polynomial models, from which one can choose those considered best for a particular purpose. It is unlikely that the same model would be selected for short gap reconstruction, for forecasting the phenomenon (with a particular time horizon), or for gaining physical insight. This approach is possible with EPR because it does not have a rigid structure, but allows a multi-structure strategy with multiple performances where each different structure has its own advantages for a specific modelling goal.

EPR is tested and demonstrated on a UK environmental case study analysing thermal behaviour of a river. Air temperature (input) and water temperature (output) data were available, but the data series had several gaps of different duration. Therefore, several models were constructed to reconstruct (infill) data [Bennis et al., 1997], obtain a model for on-line forecasting; and discover some new knowledge about the dynamics of the heat transfer process over a short time scale. In summary, the case study contains all the features that typify the analysis of an environmental phenomenon.

2. THE EVOLUTIONARY POLYNOMIAL REGRESSION

2.1 A general portrait

EPR is a data-driven hybrid method for a multi-model approach based on evolutionary computing. A general EPR expression may be given as

$$\hat{y} = \sum_{j=1}^m f(\mathbf{X}, a_j) + a_o \quad (1)$$

where \hat{y} is the estimated output of the system/process; a_j is a constant value; f is a function constructed by the process; \mathbf{X} is the matrix of input variables; and m is the length (number of terms) of the polynomial expression (bias excluded) [Giustolisi and Savic, 2004a].

The general functional structure represented by $f(\mathbf{X}, a_j)$ is constructed from elementary functions by EPR which uses a Genetic Algorithm (GA) strategy [Goldberg, 1989]. The GA is employed to select the useful input vectors from \mathbf{X} to be combined. The building blocks (elements) of the $f(\mathbf{X}, a_j)$ structure are defined by the user based on physical process understanding. While the selection of feasible structures to be combined is done through an evolutionary process, the parameters a_j are estimated by the Least Square (LS) method.

The process starts with a GA searching through the space of user defined exponents, which must include the value of zero, thus allowing a simple expression to be generated by discarding unnecessary components of \mathbf{X} . The next step consists of determining the a_j values by simple LS.

The LS is performed in an original way, by searching for only positive values. This is because negative terms usually have a poor physical meaning, as they simply balance positive terms returning a better description of noise. Neglecting negative terms constrains the search space thus gaining computational efficiency without losing accuracy. Moreover, we can hypothesize that sometimes the pressure to find parsimonious expressions could improve the search of physically based equations. In this way, EPR returns models that are probably less appropriate to on-line forecasting, but have the advantage of giving physical insight, consistent with the multi-model concept.

2.2 Some properties of EPR

EPR is a technique for data-driven modelling, successfully tested on environmental problems [Giustolisi et al., 2004a; Giustolisi and Savic, 2004b]. The combination of the GA for finding feasible structures and the LS for training a few positive constants of those structures implies some advantages. In particular, the GA allows a global exploration of the error surface thanks to specifically defined objective (cost) functions. Through such functions we can set criteria for the search: (a) avoiding the overfitting of models to training data; (b) pushing the methods towards simple structures; and (c) avoiding superfluous terms representative of the noise in data. EPR shows robustness and in every situation can get a model truly representative of data.

The use of LS for evaluation of positive constant values a_j is not compromised in working with series containing insufficient data. Indeed, LS performed on short-length expressions shows that

long time series are not necessary for proper evaluation of those constants.

In this scenario, the interesting feature of EPR is in the possibility of getting more than one model. Each of these models can be used for a specific purpose. For instance, we can get a model for a short time forecasting, another one for long time forecasting, another one for simulation, etc. Each different model can be trained according to specific cost functions.

A further feature of EPR is the high level of interactivity between the user and the methodology. The former can use physical insight to make hypotheses on the elements in the function $f(\mathbf{X}, a_j)$ and on its structure, see Eq. (1). Choosing the proper objective (cost) function and assuming pre-chosen elements in Eq. (1) (external information), and working with dimensional information enables refinement of final models [Giustolisi et al., 2004].

Finally, the best models are chosen on the basis of their performances on a test set of unseen data. For this purpose, the data set is split in two subsets: (1) the subset used for building models, named training set, and (2) the subset used for testing the model, named test set. It is important to emphasise that the test set is never used in the phase of model construction, thereby allowing us to evaluate the generalisation capacities of each model. Thus, an unbiased performance indicator is obtained on real capability of the models. Nevertheless, a bootstrap technique can also be applied to increase the robustness of model evaluations.

3. THE CASE STUDY

3.1 The River Barle

The River Barle is the main tributary of the upper River Exe. It is located in a rural zone of South-west England [Webb et al., 2003]. Our data collection consists of two years of hourly air (input) and water temperature samples (output). Each sample is referred to a window of 6 hours of a solar day covering the periods: 1-6; 7-12; 13-18; 19-24. We reasonably assume that the chosen windows are representative of the thermal dynamics at a day scale. Both air and water temperatures show two main periodic components: the annual and the daily cycles [Webb et al. 2003].

Further details about data and sampling location can be found on Webb et al. [2003].

3.2 Background to data

Before starting the modelling phase, we divided data into two subsets (training and test) each made up of 1460 samples, covering a solar year and affected by missing samples [Table 1].

Gaps in data are randomly distributed. While the longest gap is located in the test set, the 124-sample gap corresponds to 31 missing days. It should also be noted that no pre-processing was executed on gaps prior to passing the data to EPR.

A comprehensive examination of the data confirmed that the quality of samples is sufficiently good. There are neither occasional nor systematic errors which could affect modelling.

Table 1. Features of gaps contained in data.

Length of gaps in 6-hour samples	Length of gaps in hourly samples	Number of gaps
1	6	14
2	12	2
3	18	1
28	168	2
29	174	1
63	378	1
65	390	1
124	744	1

4. THE MODELLING PHASE

4.1 The strategy

The modelling phase was done as follows:

- The structure of Eq. (1) is assumed polynomial.
- Each monomial term of Eq. (1) consists of elements from \mathbf{X} raised to pre-specified power values.
- No hypotheses are made about a_0 , besides its positive sign.
- The assumed range of possible exponents of terms from \mathbf{X} is (0; 0.5; 1; 2).
- The maximum length of polynomial structures was assumed to be 5 terms plus bias.
- 7 objective (cost) functions have been used.
- The LS search is performed for positive coefficients only (negative ones are *a-priori* discarded).

- Data were scaled between 0 and 1; the outputs were rescaled before being listed.

Each objective (cost) function is based on the use of the Sum of Squared Error (SSE); the differences among them relate to the way the SSE is computed [Giustolisi and Savic, 2004a]. In summary the following cost functions were used:

- Soft Cross Validation, SCV from SSE evaluated on the whole training set.
- Rigid Cross Validation, RCV from SSE evaluated on 50% of samples of the training set.
- Control of Variance, CVP of each Parameter a_j .
- Penalization of Complex Structure, PCS.
- Penalization of Variance, PV.
- Traditional SSE evaluation, SSE.
- Control on Variance CVT of each monomial term, contained in the polynomial expression.

Details about cost functions can be found in Giustolisi and Savic [2004a]. The method of modelling ensured that the complex models with large monomial terms focussed on describing the physical process, rather than the background noise.

The presence of seven objective (cost) functions enabled a more robust multi-modelling approach, in which models can be selected according to different, appropriate objective (cost) functions to represent the most robust choices. Thus, each model has its specific utility according to the purposes previously described in the multi-model scenario. In our case study, when the search was constrained to polynomial expressions only made by 1 term plus bias, the same equation was always obtained for every cost function. By constraining the search to an expression of 2 terms plus bias, similar models were found and in some cases (e.g. PCS, PV, CVT) they were the same. Furthermore, for the same cost function, we can observe that EPR does not select more terms than it actually needs. For example, if the maximum polynomial length was set to 5, EPR could return an expression of 2 terms, because it does not consider longer expressions better than 2. Therefore, assuming a cost function, it is not unusual that after a well defined polynomial length, EPR goes on selecting the same model as optimum. On these bases, we can make a robust choice of the model. A model selected by different cost functions, or preferred to a longer expression by the same cost function, is likely to be a good model. Among

those models assumed as robust choices, the best is selected to suit the purpose. To infill gaps in data, models would be selected according to the gap length. If physical insight was required, selection would be based on those models that were easily interpretable, i.e. with a clear physical meaning.

The choice of the best models for our purpose is made on the basis of their performances on the test set. We use the Coefficient of Determination, CoD, as the main performance indicator,

$$\text{CoD} = 1 - \frac{N-1}{N} \frac{\sum_N (\bar{W} - W_{\text{exp}})^2}{\sum_N (W_{\text{exp}} - \text{mean}(W_{\text{exp}}))^2} \quad (2)$$

where N represents the number of samples, W_{exp} represents the measured water temperature, \bar{W} represents the value of water temperature returned by the model.

Furthermore, a bootstrap procedure [Efron, 1979] was used on the test set data, rather than taking a simple value of CoD. Thus, the bootstrap CoD is an average value of the CoDs evaluated by re-sampling data 1000 times from the test set. To ensure improved evaluation of the models, the standard deviation of the CoD values, reported as percentage of the average value are used. The bootstrap is particularly helpful for infilling missing data, since there are no data for comparison with model results.

4.2 EPR results

EPR returned 13 different models: we selected 3 models among them as optimal. The three selected models are,

$$W_t = 0.30574 \cdot A_t^{0.5} \cdot A_{t-1} + 0.50436 \cdot W_{t-1} + 0.31731 \cdot W_{t-4} \cdot A_t^{0.5} + 0.013418 \quad (3)$$

$$W_t = 0.078319 \cdot W_{t-4} + 0.23946 \cdot W_{t-3}^{0.5} \cdot W_{t-4}^{0.5} \cdot A_{t-1}^{0.5} + 0.49433 \cdot W_{t-1} + 0.31486 \cdot A_t \cdot A_{t-1}^{0.5} + 0.0047945 \quad (4)$$

$$W_t = 1.0073 \cdot W_{t-1}^{0.5} \cdot A_t^{0.5} \quad (5)$$

where the subscript t stands for time, in terms of 6-hour sampling rate and A refers to air temperature.

Eq. (3) is the best performing for 1-step-ahead prediction. Eq. (4) is the best working in 2-step-ahead, 4-step-ahead and 6-step-ahead prediction (one step corresponds to 6 hours). Eq. (5) was

chosen as the best working in simulation because of its more likely physical behaviour, than the best CoD-best-working model for simulation, which generates unlikely overestimated values for peak zones, because it does not contain W terms. Indeed, it does not take into consideration the effects

related to the thermal inertia of the stream, through the thermal capacity of water.

Table 2 shows the performances of the resulting models, in term of average CoD and percentage standard deviation.

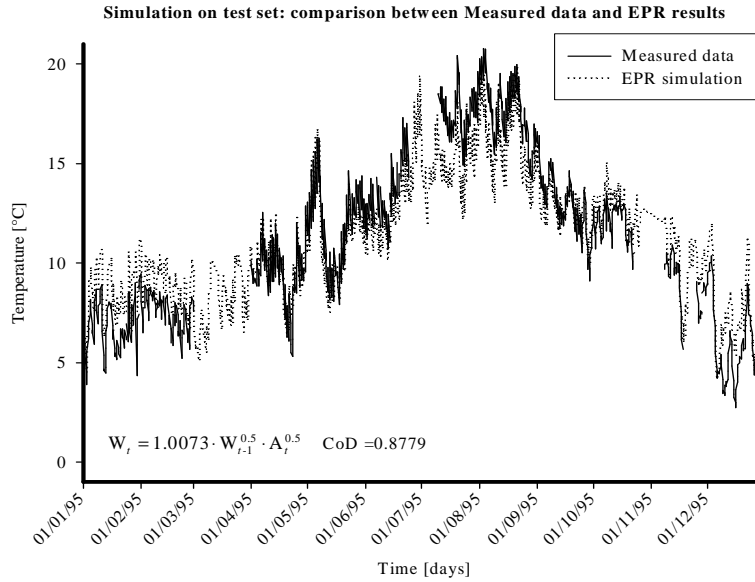


Figure 1. Comparison between measured data and EPR simulated data.

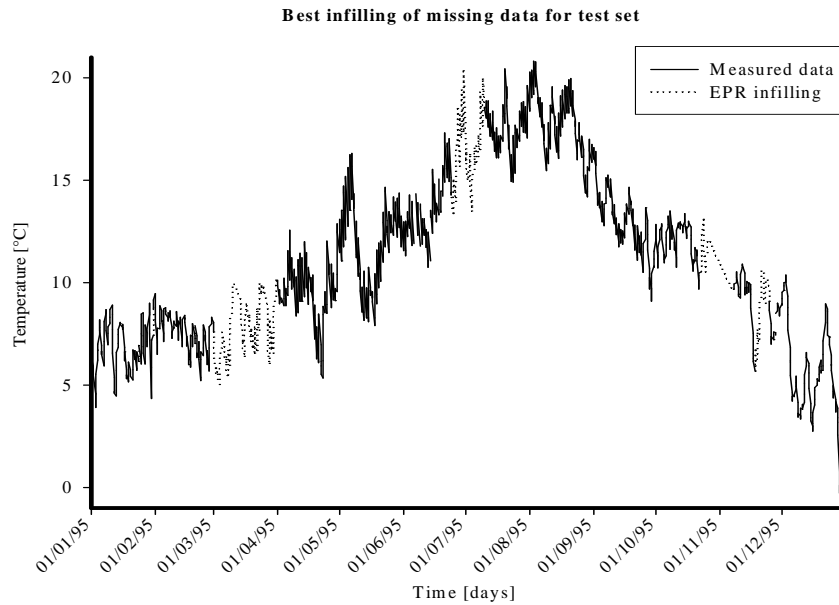


Figure 2. Best infilling of missing data.

Table 2. Statistics on model performances.

EPR model	Evaluation on	Average CoD	Standard deviation %
(3)	1-step-ahead	0.984	0.101

(4)	2-step-ahead	0.971	0.173
	4-step-ahead	0.958	0.256
	6-step-ahead	0.934	0.400
(5)	Simulation	0.878	0.412

The three models were obtained by using the following cost functions:

- CVT for Eq. (3).
- SCV for Eq. (4).
- RCV for Eq. (5).

Note that all the cost functions produced models with similar performances. We consider similar performances as strong indicators of the robustness of EPR methodology.

4.3 Comments on EPR results

All models found have a simple structure, enable the gaps in data to be reconstructed and are good at on-line forecasting and simulation. Such structures can allow a physical interpretation. In particular, Eq. (3) and Eq. (4) suggest a strong link between the output water temperature at time t and the water temperature at the time $t-1$. This interpretation is confirmed by the presence in both equations of the term W_{t-1} , multiplied by the higher coefficient in the expression. This occurred very frequently in all models. Another frequent term is the product between A_t and A_{t-1} , indicating a likely effect of the air temperature at time t and $t-1$ on water temperature. Webb confirmed this by physical observations and with a different approach to data analysis [Webb et al., 2003]. Further terms contained by models are considered of uncertain origin, and more likely associated with the background noise in data. The simulation model, whose time plot is represented in Figure 1, has a very compact unbiased expression. This is due to the similar shape, on average, of the curves representing the time plot of air temperature and water temperature. Previous studies [Webb et al., 2003; Mohseni and Stefan, 1999] underline the quasi-linear relationship between water and air temperatures, which is confirmed by our simulation model. Neglecting the stochastic information from measured data, the simulation emphasizes the quasi-linear relationship, and the W_{t-1} component seems to explain the non-linear behaviour that occurs in particular ranges of temperature [Webb et al., 2003]. Finally, in Figure 2, we can see the best infilling of data in the test set. We infilled using Eq. (4) for short and medium-size gaps, and Eq. (5) for long-size gaps. Maximum care was taken to ensure that the reconstructed values were physically possible. However, the missing samples ranging between the 23/10/95 and the 08/11/95 (right side of Figure 2), are not well reconstructed, because of the same size gap in the input data series. Therefore we linearly ap-

proximated the missing air temperatures, thereby obtaining reproduction of water temperatures.

5. CONCLUSIONS

EPR results for the case study show the effectiveness of the multi-model approach in dealing with environmental problems. We proved the ability of EPR to get parsimonious and efficient models, which can be flexibly adapted to an accurate on-line forecasting and simulation. The case study confirmed the real capabilities of the multi-model approach enabled by EPR. Additionally, the multi-model EPR strategy not only gave a good physical insight of the phenomenon, but also helped fill missing data, resulting in models that forecast the water temperature. The analysis of similar models returned by different objective (cost) functions ensured a robust choice of the best models. The cost functions were of general type, and not designed specifically for this case study, suggesting that EPR can be used without much customising for a particular problem.

6. REFERENCES

- Bennis, S., Berrada, F. and Kang, N., Improving single-variable and multivariable techniques for estimating missing hydrological data, *J. of Hydrology*, 191, 87–105, 1997.
- Davidson, J.W., Savic, D.A., & Walters, G.A., Symbolic and Numerical Regression: A Hybrid Technique for Polynomial Approximators, In *Advances in Soft Computing: Soft Computing Techniques and Applications*, R. John and R. Birkenhead (eds.), Physica-Verlag, Heidelberg, pp.111-116. 1999.
- Davidson, J.W., D.A. Savic, and G.A. Walters, Symbolic and Numerical Regression: Experiments and Applications, *Information Sciences*, 150 (1/2), pp. 95-117, 2003.
- Efron, B., Bootstrap Methods. Another Look at the Jackknife, *The Ann. of Statist.*, 7:1 – 26, 1979.
- Giustolisi, O., and Savic, D.A., A Symbolic Data-driven Technique: Evolutionary Polynomial Regression, *J. of Comp. in Civil Eng.*, ASCE, in preparation, 2004a.
- Giustolisi, O., Savic, D.A., Decision Support for Water Distribution System Rehabilitation Using Evolutionary Computing, ACTUI seminar, Exeter, 2004b.
- Giustolisi, O., Savic, D.A., and Doglioni A., Data Reconstruction and Forecasting by Evolutionary Polynomial Regression, 6th Int. Conf. on Hydroinformatics, Singapore, 2004a.

- Giustolisi, O., Savic, D.A., Doglioni A., and Laucelli, D., Knowledge discovery by Evolutionary Polynomial Regression, 6th Int. Conf. on Hydroinformatics, Singapore., 2004b.
- Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison and Wesley, Reading, Mass., USA, 1989.
- Mohseni, O., and Stefan H. G., Stream temperature/air temperature relationship: a physical interpretation, *Journal of Hydrology*, Elsevier, 218, 128 – 141, 1999.
- Webb, B.W., Clack, P.D., and Walling, D.E., Water-air temperature relationships in a Devon river system and the role of flow, *Hydrological Processes*, No.17, pp.21-37, 2003.